2-1-1992

Hybrid Databases

Carol Tenopir
University of Tennessee - Knoxville

ONLINE DATABASES

BY CAROL TENOPIR

Hybrid Databases

IT USED TO be easier to categorize online databases. Bibliographic (a.k.a. indexing/abstracting) databases contained citations and other identifying information as well as subject indexing terms and sometimes abstracts; full-text (or complete) databases included the complete textual-only portions of articles or reports; directory (referral) databases provided names, addresses, and other factual information about people or organizations; and numeric databases offered statistics about all sorts of things.

Actually, the distinctions have always been somewhat fuzzy. Online systems such as DIALOG, BRS, and Mead rarely offer purely numeric data; many directory databases are actually complete texts of printed reference works; records in full-text databases also have bibliographic information.

Still, basic categories worked pretty well for major databases for a long time. Now, such categories are less relevant as the trend moves toward mixed-type or *hybrid* databases. Hybrid databases combine one or more database types, such as bibliographic/full text, bibliographic/directory, directory/full text, etc. (On CD-ROM they may also incorporate other elements, such as full text/sound.)

Although DIALOG now lists only about 30 mixed-type databases out of its total 400 databases, the percentage across all online systems is probably higher. *Online Database Selection: A User's Guide to the Directory of Online Databases* (Cuadra/Elsevier, 1989) shows hybrid ("combinations") databases to be already the second most common type (see figure). That percentage will surely increase in 1992. In this column, I'll concentrate on bibliographic/full text, a growing trend in database enhancement. The mixture offers more for your money and is making some old databases take on a new look.

Carol Tenopir is Associate Professor at the School of Library and Information Studies, University of Hawaii at Manoa, Honolulu

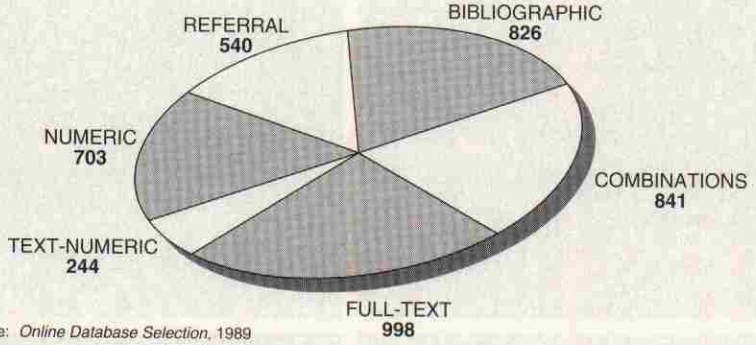
Mixing record types

Hybrid databases mix different types of records within a single searchable file. They are not the same as *linked* files such as Magazine Index or Trade & Industry Index. In linked-file databases, only the bibliographic records are searchable, but after a bibliographic search is completed, some full-text documents can be displayed because they are linked to bibliographic records. Full text in these databases is really just a document delivery solution—if you want to search words in the text you must switch to sister full-text files such as Magazine ASAP or Trade & Industry ASAP.

Hybrid databases, on the other hand, include some records that are just bibliographic and other records that have full texts. When you search, you are searching in a combined inverted index of words from both the bibliographic-only records and from the full-text records. Your output will intersperse both types of records. This causes some necessary changes in search strategy, which I'll discuss in a minute.

Some of the most frequently searched online bibliographic databases have recently gone hybrid, and some others are being created that way from the start. These are not bibliographic databases in transition to completely full text—they will stay a combination of the two types, although the full text percentage is growing in all cases. Three of the most well-known and most-searched hybrid databases are: 1) ABI/INFORM from UMI/Data Courier; 2) PTS PROMPT from Predicasts; and 3) Health Periodicals Database from Information Access Company (IAC).

Number of Online Databases by Type



Source: *Online Database Selection*, 1989

ANNABELLE CARTER

ABI/INFORM

ABI/INFORM regularly appears on the lists of online databases that are most often used in libraries. A standard for bibliographic searching of business management literature, ABI goes back to 1971 and now includes over half a million bibliographic records taken from 800 business journal titles. It is available on many different online systems, including DIALOG, BRS, NEXIS, Data-Star, and others. It is also available on CD-ROM as part of UMI's ProQuest product line. A full-text image CD-ROM system can be used in conjunction with the CD-ROM product to provide document delivery of articles from many of the ABI titles.

Until 1991, ABI/INFORM was just a well-known and well-respected indexing/abstracting product. Each record included full bibliographic data, plus lengthy informative abstracts and in-depth indexing with controlled vocabulary descriptors. Starting in August 1991, UMI/Data Courier began adding full texts of

some records to the DIALOG version of ABI. Full-text coverage on DIALOG now goes back to January 1991.

Of the current 800 ABI journal titles, articles from 105 are now available in full text. In 1992, the number of titles will grow to 250. Full-text records represent 20 percent of each month's additions to the file, and that percent will grow to 40 percent sometime in 1992.

PTS PROMPT

PTS PROMPT has been produced by Predicasts since the early 1970s. (Late last year, Predicasts was purchased by Ziff Communications, bringing it into the same family as Information Access Company.) It is the most popular of Predicasts' databases, covering a wide range of product, company, marketing, and other business information from over 1000 source documents.

For many years, PTS PROMPT was a bibliographic database whose records included descriptors and abstracts. In December 1989, full-text records from over 200 selected journal titles were added.

PTS PROMPT is a huge file—on DIALOG it includes 2.5 million records in total. Only a fraction of those are now full text, but in 1991 approximately 50 percent of the records input included the full text. Currently, about 333,000 records are full-text articles from journals, with some additional short product announcements taken from press releases.

According to Predicasts, although "the majority of titles are abstracted only, we take a few important titles for full text." The journals that are available in full text are mostly titles of major interest that are indexed nearly "cover to cover." These typically contain more articles per issue than the average journal title, so close to 50 percent of all records input last year were full text.

Health Periodicals Database

Unlike ABI/INFORM and PTS PROMPT, Health Periodicals Database is a relatively new database that was hybrid from the start. It is aimed at the layperson, in addition to the medical professional and industry researcher. It covers all aspects of health, medicine, fitness, and nutrition.

Health Periodicals is an interesting hybrid because it contains three distinct types of records: full text, indexing only, and abstracts. All three types include basic bibliographic information and controlled vocabulary indexing terms.

Most of the abstracts are actually "consumer summaries," summaries written for the layperson of technical articles from 100 professional medical journals. There are indexing records for an additional 140 titles and full text for approximately 120 professional and consumer health publications. Bibliographic records of any health-related article from IAC's other files are also placed into Health Periodicals Database. These all have indexing, and some have abstracts as well.

As of December 1991, there were approximately 274,000 records total in the database. Over 72,000 of these include the full text, with an additional 27,000 with abstracts only. (Some full-text records also have abstracts.)

Search strategies

One drawback with hybrid databases is the difficulty in formulating the most effective search strategy. Best search strategies for bibliographic databases and those for full-text databases often differ. The same free-text strategies you've been using all along will now bias your results to the full-text records in the files. Just because there are more words in a full-text record, your odds of retrieving the full-text records are greater.

Even if you are a long-time searcher of ABI/INFORM or PROMPT, it is time to rethink search techniques. In any full-text/bibliographic hybrid database the AND operator can cause difficulties. AND works perfectly well for bibliographic searching and is an ingrained habit for many searchers. Full-text searchers use the AND usually only for searching between fields—subject words anywhere in the text AND a certain publication date or author, for example. When lengthy full-text records are mixed into a bibliographic database, continued use of the AND for free-text subject searching will often result in too many false drops.

Instead, replace AND with a proximity operator recommended for full text. On DIALOG that is the (S) operator, meaning same grammatical paragraph (subfield) in full texts or same field in bibliographic records. [The (F) operator shouldn't be used even though it also means "same field" because it treats the full text as a single field rather than as grammatical subfield paragraphs.] DIALOG's (N) (near) operator with a number like (10N) or (20N) also works well. On BRS or Data-Star, use the SAME paragraph operator or the WITHin the same sentence operator.

Truncation should now be used with caution to avoid too many full-text hits. Specified stem truncation is the strategy of choice with both full-text and hybrid databases. Instead of "comput?" in ABI/INFORM on DIALOG, it is better to search "comput???" (or "comput$3" on BRS and Data-Star) if you really only need computer, computers, or computing, and not computations, computational, etc.

When viewing records, take advantage of KWIC features to view only those portions of the record that contain your search terms. Actually, especially in a hybrid database, bibliographic citation plus KWIC is the best ("TYPE set#/3,K" in DIALOG).

If you are getting too many false drops from full text, you can always eliminate the text from your search by restricting search terms to title, abstract, and descriptor words ("/ti,ab,de" on DIALOG). Data-Star makes this even easier in Health Periodicals by putting together all of the subject fields except text in a "keyword superlabel" (".KW").

Hybrid databases often offer the best of both worlds for search strategy. ABI, PROMPT, and Health Periodicals all include controlled vocabulary indexing in their full-text as well as bibliographic records, so even with full-text records you can take advantage of thesauri and the precision that comes with descriptors. Using descriptors for highly posted or potentially ambiguous terms ANDed with free-text words from titles, abstracts, or texts for jargon concepts is often the most effective combination strategy.

More bibliographic databases will be adding full-text records for those journals for which they can secure electronic republication rights. The proportion of full-text records in all hybrid databases will undoubtedly grow. As long as searchers are aware of the best ways to search these mixed-up databases, hybrid databases help solve document identification *and* document delivery problems.